



Empirical likelihood confidence bands for mean functions of recurrent events with competing risks and a terminal event

Jean-Yves Dauxois, Alexis Flesch, Davit Varron

► To cite this version:

Jean-Yves Dauxois, Alexis Flesch, Davit Varron. Empirical likelihood confidence bands for mean functions of recurrent events with competing risks and a terminal event. 2011. hal-00602699

HAL Id: hal-00602699

<https://hal.science/hal-00602699>

Preprint submitted on 23 Jun 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Empirical likelihood confidence bands for mean functions of recurrent events with competing risks and a terminal event

Jean-Yves Dauxois ^{*} Alexis Flesch [†] Davit Varron [‡]

June 23, 2011

Abstract

In this paper, we study recurrent events with competing risks in the presence of a terminal event and a censorship. We focus our attention on the mean functions which give the mean number of events of a specific type that have occurred up to a time t . Using empirical likelihood ratio techniques, we are able to build confidence bands for these functions. We have a data set of nosocomial infections in an intensive care unit of a french hospital. For each patient, we know if and when he caught an infection, what infection it was (septicemia, urinary tract infection...), if and when he died and when he left the hospital. Our model fits this context and will be used to build confidence bands for one type of nosocomial infection and even a confidence tube for two types.

Keywords: Censoring, competing risks, confidence bands, empirical likelihood, empirical processes, recurrent events, terminal event

1 Introduction

In this paper we are considering a data set on nosocomial infections contracted by more than 7867 patients in an intensive care unit of a French hospital over a decade. This data set has already been introduced and studied by Dauxois & Sencey (2009). For each patient, we know if and when he contracted a nosocomial infection and what type of infection it was: pneumonia, septicemia, urinary tract infection, etc... Each type of infection can affect the same patient several times. We also know if and when he died, and if not, when he left the hospital. This leads us to working on recurrent events with competing risks under random censorship and with a terminal event. The aim of this paper is to build confidence bands for the mean number of infections of one or more types over time. This will be achieved using empirical likelihood and empirical processes methods.

Recurrent events models are useful in many fields, like in Social Science for recurrent periods of unemployment, in Reliability for recurrent occurrences of failure on the same device or also in Biostatistics as in our case study of nosocomial infections, etc...

Many statistical methods have been used over the last decades to study recurrent events based on Markov models, Martingale theory or Poisson processes. Andersen *et al.* (1993) give a tour of these methods. Lawless & Nadeau (1995) and Lawless (1995) considered the mean function, which gives the mean number of events that have occurred up to a time t . Wang & Wells (1998), Lin *et al.* (1999), Lin & Ying (2001), Cai & Schaubel (2004a) and more recently Du (2009) focused on the gap time distribution. More recently, Cook & Lawless (2007) issued a book on the analysis of recurrent events.

In some situations, a terminal event may stop the recurrent event process, as in our data set where the death of the patient is clearly dependent on the infections he suffered from, or when

^{*}Laboratoire de Mathématiques de Besançon, UMR CNRS 6623, Université de Franche-Comté, 16 route de Gray, 25030 Besançon cedex, France (jean-yves.dauxois@univ-fcomte.fr).

[†]Laboratoire de Mathématiques de Besançon, UMR CNRS 6623, Université de Franche-Comté, 16 route de Gray, 25030 Besançon cedex, France (alexis.flesch@univ-fcomte.fr).

[‡]Laboratoire de Mathématiques de Besançon, UMR CNRS 6623, Université de Franche-Comté, 16 route de Gray, 25030 Besançon cedex, France (dvarron@univ-fcomte.fr).

a patient leaves the hospital. The end of observation can also be due to an independent right censoring mechanism, specifically, as in our case, when the study comes to an end. Note that in other contexts, other causes of independent censoring can be observed like in the case of lost to follow up. Papers considering recurrent events with a terminal event and independent censoring are among others Cook *et al.* (2009) and Cai & Schaubel (2004b).

Finally, there are some situations where the observed events are of multiple types. It is the case in our data set since several types of nosocomial infections can affect a patient: pneumonia, septicemia, urinary tract infection, etc... In the same manner, the failure of a device can be caused by multiple mechanical parts. As an example, the engine, the gearbox or the tires can be a cause of failure for cars. This is also true for recurrent unemployment: one can loose his job for numerous different reasons. It is thus interesting to consider competing risks in a setup of recurrent events, as seen in Dauxois & Sencey (2009).

Empirical likelihood was first introduced by Thomas & Grunkemeier (1975) in a setup of survival analysis. Owen (1990) generalized their concept to obtain confidence regions for the mean of an i.i.d. sample with finite variance. One of the advantages of this approach is that one doesn't need to estimate the variance to build the region. Moreover, the shape of those regions strongly depend on the geometry of the data whereas classical central limit theorem gives ellipse-shaped confidence regions. Since Owen (1990), a lot of results on empirical likelihood have been published, and a few concern survival analysis. For example, Adimari (1997) and Ren (2001) used empirical likelihood to study the mean under random censorship. Owen (2001) issued a book on empirical likelihood, and, more recently, Hjort *et al.* (2004) gave a generalization of Owen's result which will be used in this article.

Our aim in this paper is to build confidence bands for the mean functions of a multiple-type recurrent events process with terminal event and under right censoring. In order to do so, we use results from Hjort *et al.* (2004) and empirical processes methods, available for example in Pollard (1990), Pakes & Pollard (1989) and Biliias *et al.* (1997).

In section 2, we introduce the model and the empirical likelihood ratio for a specific mean function of a recurrent event process. The asymptotic distribution of the empirical likelihood ratio is obtained in section 3 and then we generalize this result to the empirical likelihood ratio for two specific mean functions. Finally, in section 4, we apply our results to the data set of nosocomial infections and obtain confidence bands for the mean functions.

2 Notations

For seek of simplicity, we shall suppose that only two types of events are observed, as the generalization is straightforward. For $j = 1, 2$, we write $N_j^*(t)$ the total number of events of type j that have occurred up to time t . We suppose that the counting processes N_j^* are almost surely bounded by a constant B . We also suppose the presence of a terminal event D *a priori* dependent on the N_j^* 's after which the counting processes can't jump. Finally, we suppose that the observation of the processes is subject to a censorship C which is independant on the N_j^* 's and D .

Write $X = D \wedge C$ the time after which the process doesn't jump anymore and $\delta = I(D \leq C)$, where $I(\cdot)$ denotes the indicator function. This last random variable informs us on whether it was the terminating event that stopped the process or the censorship. If we write $Y(t) = I(X \geq t)$ and $N_j(t) = N_j^*(t \wedge C)$ for $t \geq 0$ and $j = 1, 2$, then the observed data are i.i.d. replicates $(N_{i,j}(t), X_i, \delta_i)$, for $i = 1, \dots, n$ and $j = 1, 2$, of $(N_j(t), X, \delta)$, for $t \in [0, \tau]$ where τ is a fixed constant chosen so that $\mathbb{P}(C > \tau)\mathbb{P}(X > \tau) > 0$.

In order to build confidence regions for the mean functions $t \mapsto \mu_j(t) = \mathbb{E}(N_j^*(t))$, we use the estimators $\hat{\mu}_j(t)$ from Dauxois & Sencey (2009):

$$\hat{\mu}_j(t) = \frac{1}{n} \sum_{i=1}^n \int_0^t \frac{\hat{S}(u^-)}{\bar{Y}(u)} dN_{i,j}(u),$$

where \hat{S} is the Kaplan-Meier estimator of S , the survival function of D (see, for example, Andersen *et al.* (1993) for a definition of \hat{S}) and $\bar{Y}(t) = \frac{1}{n} \sum_{i=1}^n Y_i(t)$. We then apply empirical likelihood methods. Formally, write for all t in $[0, \tau]$:

$$\text{EL}_n(\mu_j, t) := \max \left\{ \prod_{i=1}^n np_i, \hat{\mu}_{j,p}(t) = \mu_j(t), p_i \geq 0 \text{ and } \sum_{i=1}^n p_i = 1 \right\},$$

where:

$$\begin{aligned} \hat{\mu}_{j,p}(t) &:= \int_0^t \hat{S}(u^-) d\hat{R}_{j,p}(u), \\ \hat{R}_{j,p}(t) &:= \sum_{i=1}^n \int_0^t p_i \frac{dN_{i,j}(u)}{\bar{Y}(u)}. \end{aligned}$$

Straight calculus shows that:

$$\hat{\mu}_{j,p}(t) = \mu_j(t) \iff \sum_{i=1}^n p_i \left(\int_0^t \frac{\hat{S}(u^-)}{\bar{Y}(u)} dN_{i,j}(u) - \mu_j(t) \right) = 0.$$

Following notations in Hjort *et al.* (2004), we write:

$$\hat{h}(u) = \frac{\hat{S}(u^-)}{\bar{Y}(u)} \quad \text{and} \quad m_{i,j}(N_{i,j}, \mu_j, \hat{h}, t) = \int_0^t \hat{h}(u) dN_{i,j}(u) - \mu_j(t).$$

EL_n can now be rewritten as in Hjort *et al.* (2004):

$$\text{EL}_n(\mu_j, \hat{h}, t) = \max \left\{ \prod_{i=1}^n np_i, \sum_{i=1}^n p_i m_{i,j}(N_{i,j}, \mu_j, \hat{h}, t) = 0, p_i \geq 0 \text{ and } \sum_{i=1}^n p_i = 1 \right\}.$$

Finally, define:

$$M_{n,j}(\mu_j, \hat{h}, t) = \frac{1}{n} \sum_{i=1}^n m_{i,j}(N_{i,j}, \mu_j, \hat{h}, t) \quad \text{and} \quad S_{n,j}(\mu_j, \hat{h}, t) = \frac{1}{n} \sum_{i=1}^n m_{i,j}(N_{i,j}, \mu_j, \hat{h}, t)^2.$$

When no confusion is possible, we will denote by $m_{i,j}(t)$ and $\hat{m}_{i,j}(t)$ the functions:

$$\begin{aligned} m_{i,j}(t) &= m_{i,j}(N_{i,j}, \mu_j, h, t) = \int_0^t \frac{S(u^-)}{\mathbb{E}Y(u)} dN_{i,j}(u) - \mu_j(t), \\ \hat{m}_{i,j}(t) &= m_{i,j}(N_{i,j}, \mu_j, \hat{h}, t) = \int_0^t \hat{h}(u) dN_{i,j}(u) - \mu_j(t). \end{aligned}$$

One can easily show that $\mathbb{E} m_{i,j}(t) = 0$ for all $t \in [0, \tau]$.

3 Main result

3.1 Convergence of the margins

We will need the following result which is easily obtained from Theorem 2 of Dauxois & Sencey (2009):

Theorem 1. *For $j = 1, 2$, as $n \rightarrow \infty$,*

$$\sqrt{n}(\hat{\mu}_j(\cdot) - \mu_j(\cdot)) \xrightarrow{\mathcal{D}} G_j(\cdot),$$

in the Skorohod space $D[0, \tau]$, where $G_j(\cdot)$ is a mean-zero Gaussian process.

We can now state our main result: under a classical variance condition we establish the weak convergence of the empirical likelihood ratio.

Theorem 2. Suppose that both $V_j(t) := \text{Var } m_{i,j}(t)$ and $\mathbb{E}|m_{i,j}(t)|$ are bounded away from zero and infinity on $[\tau_1, \tau]$, for some real constant $\tau_1 > 0$. Then:

$$-2 \log EL_n(\mu_j, \hat{h}, \cdot) \xrightarrow{D} G_j^2(\cdot) V_j^{-1}(\cdot), \quad \text{when } n \rightarrow +\infty,$$

in the Skorohod space $D[\tau_1, \tau]$.

Proof. This result is mainly based on the section 2.3 in Hjort *et al.* (2004). It is then sufficient to show that their conditions (A0*) to (A3*) are true, that is:

$$(A0^*) : \mathbb{P}(\exists t \in [\tau_1, \tau] : \text{EL}_n(\mu_j(t), \hat{h}, t) = 0) \longrightarrow 0, \text{ as } n \rightarrow \infty.$$

$$(A1^*) : \sqrt{n} M_{n,j}(\mu_j, \hat{h}, \cdot) \xrightarrow{D} G(\cdot), \text{ as } n \rightarrow \infty.$$

$$(A2^*) : \sup_{t \in [\tau_1, \tau]} |S_{n,j}(\mu_j, \hat{h}, t) - V_j(t)| \xrightarrow{\mathbb{P}} 0, \text{ as } n \rightarrow \infty.$$

$$(A3^*) : \max_{1 \leq i \leq n} \|m_{i,j}(N_i, \mu_j, \hat{h}, t)\|_\infty = o_p(n^{1/2}),$$

where $\|\cdot\|_\infty$ denotes the supremum norm on $\mathcal{D}[\tau_1, \tau]$.

Proof of (A0).* It is sufficient to show that almost surely:

$$\liminf_{n \rightarrow \infty} \inf_{t \in [\tau_1, \tau]} \frac{1}{n} \sum_{i=1}^n I(\hat{m}_{i,j}(t) > 0) > 0 \quad \text{and} \quad \overline{\lim}_{n \rightarrow \infty} \sup_{t \in [\tau_1, \tau]} \frac{1}{n} \sum_{i=1}^n I(\hat{m}_{i,j}(t) > 0) < 1. \quad (1)$$

Let us first consider the left inequality. Note that we have almost surely, for all $\eta > 0$ and $t \in [\tau_1, \tau]$,

$$I(|\hat{m}_{i,j}(t) - m_{i,j}(t)| > \eta) \leq I\left(B \left\| \frac{\hat{S}(u^-)}{\bar{Y}(u)} - \frac{S(u^-)}{\mathbb{E}Y_1(u)} \right\|_\infty > \eta\right).$$

Thus, almost surely for all t in $[\tau_1, \tau]$:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n I(\hat{m}_{i,j}(t) > 0) &\geq \frac{1}{n} \sum_{i=1}^n (I(m_{i,j}(t) > \eta) - I(|\hat{m}_{i,j}(t) - m_{i,j}(t)| > \eta)) \\ &\geq \frac{1}{n} \sum_{i=1}^n I(m_{i,j}(t) > \eta) - I\left(B \left\| \frac{\hat{S}(u^-)}{\bar{Y}(u)} - \frac{S(u^-)}{\mathbb{E}Y_1(u)} \right\|_\infty > \eta\right). \end{aligned}$$

We know that $\|\mathbb{E}Y(u) - \bar{Y}(u)\|_\infty$ and $\|S(u^-) - \hat{S}(u^-)\|_\infty$ converge to 0 when $n \rightarrow +\infty$ from Glivenko-Cantelli Theorem and the uniform consistency of the Kaplan-Meier estimator (see e.g. Theorem IV.3.1 of Andersen *et al.* (1993)) respectively. Thus, the second indicator function tends to zero and it is sufficient to show that there exists $\eta > 0$ such that $\inf_{t \in [\tau_1, \tau]} \mathbb{P}(m_{i,j}(t) > \eta) > 0$. To this effect, let:

$$\eta = \frac{1}{2} \inf_{t \in [\tau_1, \tau]} \mathbb{E}(m_{i,j}(t) I(m_{i,j}(t) \geq 0)) = \frac{1}{4} \inf_{t \in [\tau_1, \tau]} \mathbb{E}|m_{i,j}(t)| > 0.$$

This last equality is a consequence of $\mathbb{E} m_{i,j}(t) = 0$ for all $t \in [\tau_1, \tau]$. Cauchy-Schwarz inequality yields:

$$\begin{aligned} \mathbb{P}(m_{i,j}(t) \geq \eta) &\geq \frac{1}{\mathbb{E}(m_{i,j}(t)^2)} \mathbb{E}^2(m_{i,j}(t) I(m_{i,j}(t) \geq \eta)) \\ &= \frac{1}{\mathbb{E}(m_{i,j}(t)^2)} (\mathbb{E}(m_{i,j}(t) I(m_{i,j}(t) \geq 0)) - |\mathbb{E}(m_{i,j}(t) I(m_{i,j}(t) \geq 0))| \\ &\quad - \mathbb{E}(m_{i,j}(t) I(m_{i,j}(t) \geq \eta))|)^2 \\ &\geq \frac{1}{\mathbb{E}(m_{i,j}(t)^2)} (2\eta - \eta)^2 \\ &= \frac{\eta^2}{\mathbb{E}(m_{i,j}(t)^2)}, \end{aligned}$$

which is, from our assumptions on $\mathbb{E}|m_{i,j}(t)|$ and $\mathbb{V}\text{ar}(m_{i,j}(t))$, bounded away from 0 for all $t \in [\tau_1, \tau]$.

To show the second inequation in (1), remark that:

$$\overline{\lim}_{n \rightarrow \infty} \sup_{t \in [\tau_1, \tau]} \frac{1}{n} \sum_{i=1}^n I(\widehat{m}_{i,j}(t) > 0) < 1 \iff \underline{\lim}_{n \rightarrow \infty} \inf_{t \in [\tau_1, \tau]} \frac{1}{n} \sum_{i=1}^n I(\widehat{m}_{i,j}(t) \leq 0) > 0,$$

and conclude by using the same arguments as previously.

Proof of (A1).* Recall that $M_{n,j}(\mu_j(t), \widehat{h}, t) = \frac{1}{n} \sum_{i=1}^n m_{i,j}(N_{i,j}, \mu_j, \widehat{h}, t)$. Then, by Theorem 1:

$$\sqrt{n} M_{n,j}(\mu_j, \widehat{h}, \cdot) = \sqrt{n} (\widehat{\mu}_j(\cdot) - \mu(\cdot)) \xrightarrow{D} G_j(\cdot), \quad \text{as } n \rightarrow +\infty.$$

Proof of (A2).* Recall that $S_{n,j}(\mu_j, \widehat{h}, t) = \frac{1}{n} \sum_{i=1}^n m_{i,j}(N_i, \mu_j, \widehat{h}, t)^2$. Hence:

$$S_{n,j}(\mu_j, \widehat{h}, t) = \frac{1}{n} \sum_{i=1}^n (A_{i,j}(t)^2 + 2A_{i,j}(t)m_{i,j}(t) + m_{i,j}(t)^2),$$

where:

$$A_{i,j}(t) = \int_0^t \left(\frac{\widehat{S}(u^-)}{\widehat{Y}(u)} - \frac{S(u^-)}{\mathbb{E}Y(u)} \right) dN_{i,j}(u).$$

• **Step 1.** We shall show that: $\|\frac{1}{n} \sum_{i=1}^n m_{i,j}(t)^2 - V_j(t)\|_\infty \xrightarrow{\mathbb{P}} 0$ as $n \rightarrow \infty$.

Write $f_{i,j}(\omega, t) := m_{i,j}(t)^2$ and $\mathcal{F}_{n\omega}^j = \{(f_{1,j}(\omega, t), \dots, f_{n,j}(\omega, t)), t \in [\tau_1, \tau]\}$. Write also:

$$g_{i,j}(\omega, t) = \int_0^t \frac{S(u^-)}{\mathbb{E}Y_1(u)} dN_{i,j}(\omega, u),$$

and

$$\mathcal{G}_{n\omega}^j = \{(g_{1,j}(\omega, t), \dots, g_{n,j}(\omega, t)), t \in [\tau_1, \tau]\}.$$

Then, almost surely:

$$\begin{aligned} |g_{i,j}(\omega, t)| &= \left| \int_0^t \frac{1}{\mathbb{P}(C \geq u)} dN_{i,j}(\omega, u) \right| \\ &\leq \frac{1}{\mathbb{P}(C \geq \tau)} \int_0^\tau dN_{i,j}(\omega, u) \\ &\leq \frac{B}{\mathbb{P}(C \geq \tau)}. \end{aligned}$$

Hence, the class $\mathcal{G}_{n\omega}^j$ is euclidean (see Pollard (1990) pp. 38 for a definition of euclidean). Indeed, as the $g_{i,j}$'s are non decreasing, the class is of pseudo-dimension 1 (see Biliias *et al.* (1997)). Conclude by using Theorem 4.8 from Pollard (1990). The same arguments apply to the class $\mathcal{H}_n^j = \{(\mu_j(t), \dots, \mu_j(t)), t \in [\tau_1, \tau]\}$, as $|\mu_j(t)| \leq B$. Now, notice that $\mathcal{F}_{n\omega}^j \subset (\mathcal{G}_{n\omega}^j - \mathcal{H}_n^j)^2$. As euclidean classes are stable under addition and multiplication (see Pakes & Pollard (1989)), $\mathcal{F}_{n\omega}^j$ is euclidean. Thus, since $m_{i,j}(t)$ is a mean zero process:

$$\sup_{t \in [\tau_1, \tau]} \left| \frac{1}{n} \sum_{i=1}^n f_{i,j}(t, \omega) - \mathbb{E}f_{i,j}(t, \cdot) \right| \xrightarrow{a.s.} 0, \quad \text{as } n \rightarrow +\infty.$$

• **Step 2.** Let us show that: $\|\frac{1}{n} \sum_{i=1}^n A_{i,j}(t)^2\|_\infty \xrightarrow{\mathbb{P}} 0$, as $n \rightarrow +\infty$.

Rewrite $A_{i,j}(t)$ as:

$$A_{i,j}(t) = \int_0^t \frac{1}{\bar{Y}(u)\mathbb{E}Y(u)} \left(\hat{S}(u^-)(\mathbb{E}Y(u) - \bar{Y}(u)) - \bar{Y}(u)(S(u^-) - \hat{S}(u^-)) \right) dN_{i,j}(u).$$

But, from our hypothesis on τ we have:

$$\inf_{t \in [\tau_1, \tau]} \mathbb{E}Y(t) = \mathbb{P}(D \geq \tau)\mathbb{P}(C \geq \tau) > 0,$$

and:

$$\bar{Y}(u) = \frac{1}{n} \sum_{i=1}^n I(X_i \geq u) \geq \bar{Y}(\tau),$$

where the last term is well-known to converge almost surely to $\mathbb{P}(D \geq \tau)\mathbb{P}(C \geq \tau) > 0$. Thus, almost surely:

$$\left\| \frac{1}{n} \sum_{i=1}^n A_{i,j}(t)^2 \right\|_{\infty} \leq \left(\frac{\|\mathbb{E}Y(u) - \bar{Y}(u)\|_{\infty} + \|S(u^-) - \hat{S}(u^-)\|_{\infty}}{\bar{Y}(\tau)\mathbb{E}Y(\tau)} \right)^2 B^2,$$

since $N_{i,j}$ is bounded by B . The above inequality and the convergence of both $\|\mathbb{E}Y(u) - \bar{Y}(u)\|_{\infty}$ and $\|S(u^-) - \hat{S}(u^-)\|_{\infty}$ yield the expected convergence of step 2.

• **Step 3.** Show that: $\|\frac{2}{n} \sum_{i=1}^n A_{i,j}(t)m_{i,j}(t)\|_{\infty} \xrightarrow{\mathbb{P}} 0$, as $n \rightarrow +\infty$.

The same arguments as in Step 2 apply to show that $\|\frac{1}{n} \sum_{i=1}^n A_{i,j}(t)\|_{\infty} \xrightarrow{\mathbb{P}} 0$. Conclude by noticing that:

$$\max_{1 \leq i \leq n} \|m_{i,j}(t)\|_{\infty} \leq \max_{1 \leq i \leq n} \left(\int_0^{\tau} \frac{S(u^-)}{\mathbb{E}Y(u)} dN_{i,j}(u) + \mu_j(\tau) \right) \leq \frac{B}{\mathbb{E}Y(\tau)} + B.$$

Proof of (A3).* First, notice that:

$$\begin{aligned} \|m_{i,j}(N_i, \mu_j, \hat{h}, t)\|_{\infty} &= \left\| \int_0^t \frac{\hat{S}(u^-)}{\bar{Y}(u)} dN_{i,j}(u) - \mu_j(t) \right\|_{\infty} \\ &\leq \int_0^{\tau} \frac{1}{\bar{Y}(u)} dN_{i,j}(u) + \mu_j(\tau) \\ &\leq \frac{B}{\bar{Y}(\tau)} + B. \end{aligned}$$

It is easily seen that this last sum of two terms is almost surely $o_p(n^{1/2})$ since $\bar{Y}(\tau)$ converges almost surely to $\mathbb{E}Y(\tau) > 0$ when $n \rightarrow +\infty$. Thus, condition (A3*) is fulfilled which completes the proof of Theorem 2. \square

3.2 Joint convergence

In this section we will state a corollary of our Theorem 2 where we consider the profile empirical likelihood ratio function for both mean functions μ_1 and μ_2 . This latter is defined as:

$$\text{EL}_n(\mu, \hat{h}, t) = \max \left\{ \prod_{i=1}^n np_i, \sum_{i=1}^n p_i m_i(N_i, \mu, \hat{h}, t) = 0, p_i \geq 0 \text{ and } \sum_{i=1}^n p_i = 1 \right\},$$

where

$$m_i(N_i, \mu, \hat{h}, t) = (m_{i,j}(N_{i,j}, \mu_j, \hat{h}, t))_{j=1,2}.$$

As before, when no confusion is possible, let us write:

$$m_i(t) = (m_{i,j}(N_{i,j}, \mu_j, h, t))_{j=1,2} = \left(\int_0^t \frac{S(u^-)}{\mathbb{E}Y(u)} dN_{i,j}(u) - \mu_j(t) \right)_{j=1,2},$$

$$\widehat{m}_i(t) = (m_{i,j}(N_{i,j}, \mu_j, \widehat{h}, t))_{j=1,2} = \left(\int_0^t \widehat{h}(u) dN_{i,j}(u) - \mu_j(t) \right)_{j=1,2}.$$

Before stating the corollary of our Theorem 2, recall the following result from Dauxois & Sencey (2009):

Theorem 3. *As $n \rightarrow \infty$,*

$$\sqrt{n} \begin{pmatrix} \widehat{\mu}_1 - \mu_1 \\ \widehat{\mu}_2 - \mu_2 \end{pmatrix} \xrightarrow{D} \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} =: G.$$

Corollary 4. *Suppose that the eigenvalues of $V(t)$, the covariance matrix of $m_i(t)$, are uniformly bounded away from zero and infinity on $[\tau_1, \tau]$ for some real constant $\tau_1 > 0$. Suppose also that $\mathbb{E}[m_i(t)]$ is uniformly bounded away from zero and infinity on the same interval (where the absolute value is taken composant-wise). Then:*

$$-2 \log EL_n(\mu, \widehat{h}, \cdot) \xrightarrow{D} G(\cdot)^t V^{-1}(\cdot) G(\cdot), \quad \text{as } n \rightarrow +\infty,$$

in the Skorohod space $D[\tau_1, \tau]$.

Proof. Although the quantities M_n and S_n from last section have to be adapted to the vectorial case, the proof is very similar to the one of Theorem 2. Let:

$$M_n(\mu, \widehat{h}, t) = \frac{1}{n} \sum_{i=1}^n m_i(N_i, \mu, \widehat{h}, t) \quad \text{and} \quad S_n(\mu, \widehat{h}, t) = \frac{1}{n} \sum_{i=1}^n m_i(N_i, \mu, \widehat{h}, t)^{\otimes 2},$$

where $x^{\otimes 2} := xx^t$. It is then easily seen that hypotheses **(A1*)** and **(A3*)**, after straightforward adaptation to the multivariate case, are verified.

(A0*): $\mathbb{P}(\exists t \in [\tau_1, \tau] : EL_n(\mu, \widehat{h}, t) = 0) \rightarrow 0$, as $n \rightarrow \infty$.

It is sufficient to show that almost surely, as $n \rightarrow \infty$, every half-space around 0 is occupied by the \widehat{m}_i 's. Formally, it is sufficient to show that, almost surely:

$$\lim_{n \rightarrow \infty} \inf_{t \in [\tau_1, \tau], \theta \in \Theta} \frac{1}{n} \sum_{i=1}^n I(\langle \widehat{m}_i(t), \theta \rangle > 0) > 0, \quad (2)$$

where Θ is the unit sphere of \mathbb{R}^2 . To show (2), remark that for $\eta > 0$, Cauchy-Schwarz's inequality yields:

$$|\langle \widehat{m}_i(t) - m_i(t), \theta \rangle| > \eta \Rightarrow \|\widehat{m}_i(t) - m_i(t)\|_2 > \eta,$$

where $\|\cdot\|_2$ denotes the euclidean norm. Thus:

$$\frac{1}{n} \sum_{i=1}^n I(\langle \widehat{m}_i(t), \theta \rangle > 0) \geq \frac{1}{n} \sum_{i=1}^n I(\langle m_i(t), \theta \rangle > \eta) - \frac{1}{n} \sum_{i=1}^n I(\|\widehat{m}_i(t) - m_i(t)\|_2 > \eta). \quad (3)$$

But, we have shown that the last term in (3) converges to 0 as $n \rightarrow \infty$ in probability. Let now $T_i(t, \theta) := \langle m_i(t), \theta \rangle$. We know from our hypotheses that $\text{Var} T_i(t, \theta) = \langle V(t)\theta, \theta \rangle$ is uniformly bounded away from 0 and infinity on $[\tau_1, \tau] \times \Theta$. It is also true for $\mathbb{E}[T_i(t, \theta)]$. We can then conclude by applying the same proof as in Theorem 2 to the random variable $T_i(t, \theta)$ instead of $m_{i,j}(t)$.

(A2*): $\sup_{t \in [\tau_1, \tau]} |S_n(\mu, \widehat{h}, t) - V(t)| \xrightarrow{\mathbb{P}} 0$.

Notice that:

$$S_n(\mu, \widehat{h}, t) = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} m_{i,1}^2 & m_{i,1}m_{i,2} \\ m_{i,2}m_{i,1} & m_{i,2}^2 \end{pmatrix},$$

where $m_{i,j}$ is short for $m_{i,j}(N_i, \mu_j, \hat{h}, t)$. We have already shown the convergence of the diagonal terms in the proof of Theorem 2. Using similar arguments one can show that, in probability:

$$\sup_{t \in [\tau_1, \tau]} \left| \frac{1}{n} \sum_{i=1}^n m_{i,1}(t) m_{i,2}(t) - \text{Cov}(m_1(t), m_2(t)) \right| \xrightarrow[n \rightarrow \infty]{} 0.$$

□

4 Application to nosocomial infections

4.1 Algorithm in dimension 1

In this section, we will apply our results to the data set of nosocomial infections mentioned in section 1. We will restrict our attention to pneumonies, septicemias and urinary tract infections, as those are the most frequent infections in the data set: for example, 373 patients experienced a total amount of 463 pneumonies during their stay. Using Theorem 2, one can build simultaneous confidence band for the mean function of each type of infection. We shall then describe the algorithm used to obtain a 95% confidence band for the mean function of one type of infection. Using the same notations as before, we have $n = 7867$ and $\max_{1 \leq i \leq n} X_i = 380$. First of all, we approximate the 95%-quantile of the limiting law by bootstrap. To this effect, we draw (Z_1^*, \dots, Z_n^*) uniformly in $\{Z_1, \dots, Z_n\}$, where $Z_i := (N_i, X_i, \delta_i)$. Then, for each day k in $\{0, \dots, 380\}$, we compute $-2 \log \text{EL}_n(\hat{\mu}(k), \hat{h}^*, k)$ and we store the supremum (over k) of those values. Finally, we repeat this 1000 times, sort the values and take the 950-*th* as our 95%-estimated quantile \hat{u}_{95} . Then, for each day k in $\llbracket 0, 380 \rrbracket$, we determine by dichotomy the convex set C_k for which we have:

$$\forall y \in C_k, -2 \log \text{EL}_n(y, \hat{h}, k) \leq \hat{u}_{95}.$$

Notice that the confidence band stays the same after the last jump: for example, no patient contracted a pneumonia after the 125-*th* day as can be seen on the Figure 1. Notice also that the confidence band is not centered around $\hat{\mu}$: unlike central limit theorem, empirical likelihood builds confidence regions whose shape strongly depends on the geometry of the data.

4.2 Algorithm in dimension 2

As pointed out in section 3.2, our result still holds for more than one type of events. We thus built a confidence tube for (μ_1, μ_2) where μ_1 is the mean number of pneumonies and μ_2 the mean number of septicemias. In this context, the number of patients who experienced either pneumonies or septicemias is 547, with 71 patients experiencing both types of infections at least one time. The total number of pneumonies is 463 (373 patients experienced it) and the total amount of septicemias is 289. For the 3D graph to be more readable, we projected it on a plane, eliminating the dimension corresponding to time. Each convex set corresponds to a day, the first day being at the bottom left corner and the last day at the upper right corner. As in dimension 1, one can notice that this confidence tube is not elliptic as would be expected had we used the central limit theorem. Once again, this is due to empirical likelihood.

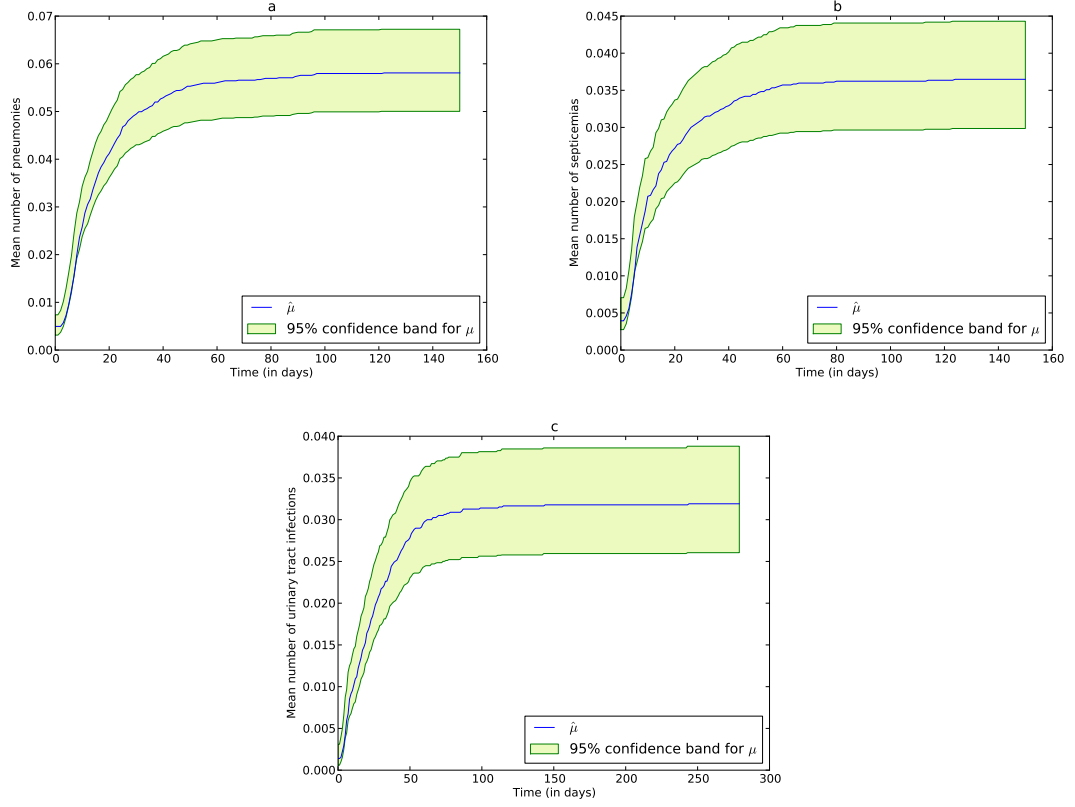


Figure 1: Nosocomial data set. Estimated specific mean function and associated 95% empirical likelihood confidence bands for a) pneumonia, b) septicemia and c) urinary tract infection.

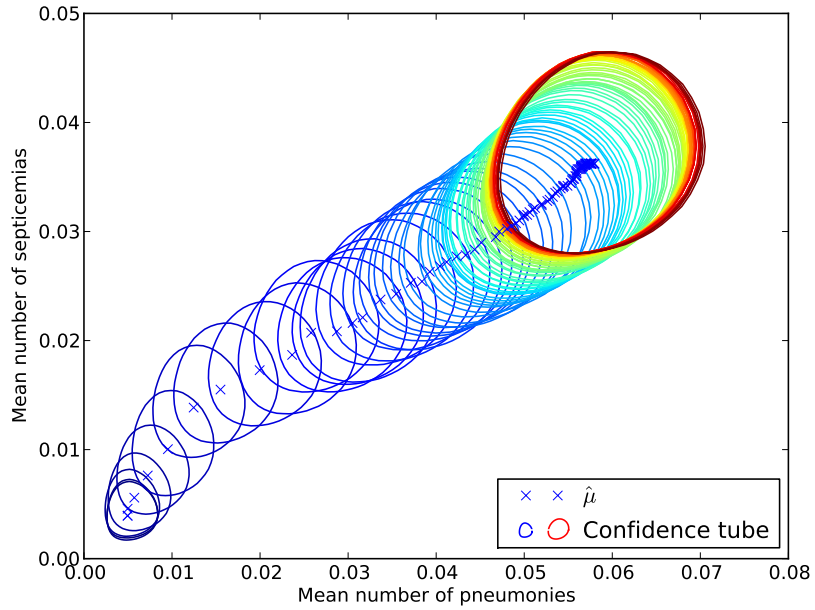


Figure 2: Nosocomial data set. Estimated 95% empirical likelihood confidence tube for the mean number of pneumonias and septicemias along with $\hat{\mu}$

References

- Adimari, G. (1997). Empirical likelihood type confidence intervals under random censorship. *Ann. Inst. Statist. Math.* **49**, 447–466.
- Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York.
- Biliyas, Y., Gu, M. & Ying, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry. *Ann. Statist.* **25**, 662–682.
- Cai, J. & Schaubel, D. E. (2004a). Analysis of recurrent event data. In *Advances in survival analysis*, vol. 23 of *Handbook of Statist.* Elsevier, Amsterdam, pp. 603–623.
- Cai, J. & Schaubel, D. E. (2004b). Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Anal.* **10**, 121–138.
- Cook, R. & Lawless (2007). *The statistical analysis of recurrent events*. Springer.
- Cook, R. J., Lawless, J. F., Lakhal-Chaieb, L. & Lee, K.-A. (2009). Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: application to skeletal complications in cancer metastatic to bone. *J. Amer. Statist. Assoc.* **104**, 60–75.
- Dauxois, J.-Y. & Sencey, S. (2009). Non-parametric tests for recurrent events under competing risks. *Scand. J. Stat.* **36**, 649–670.
- Du, P. (2009). Nonparametric modeling of the gap time in recurrent event data. *Lifetime Data Anal.* **15**, 256–277.
- Hjort, N. L., McKeague, I. W. & Van Keilegom, I. (2004). Extending the scope of empirical likelihood (preprint) .
- Hjort, N. L., McKeague, I. W. & Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37**, 1079–1111.
- Lawless, J. F. (1995). The analysis of recurrent events for multiple subjects. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **44**, pp. 487–498.
- Lawless, J. F. & Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**, 158–168.
- Lin, D. Y., Sun, W. & Ying, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* **86**, 59–70.
- Lin, D. Y. & Ying, Z. (2001). Nonparametric tests for the gap time distributions of serial events based on censored data. *Biometrics* **57**, 369–375.
- Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics* **18**, 90–120.
- Owen, A. (2001). *Empirical likelihood*. Monographs on statistics and applied probability. Chapman & Hall/CRC.
- Pakes, A. & Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57**, 1027–1057.
- Pollard, D. (1990). *Empirical processes: theory and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA.
- Ren, J.-J. (2001). Weighted empirical likelihood ratio confidence intervals for the mean with censored data. *Ann. Inst. Statist. Math.* **53**, 498–516.

- Thomas, D. R. & Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70**, 865–871.
- Wang, W. & Wells, M. T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika* **85**, 561–572.